Drug target prediction by multi-view low rank embedding

Limin Li, Menglan Cai

Abstract—Drug repositioning has been a key problem in drug development, and heterogeneous data sources are used to predict drug-target interactions by different approaches. However, most of studies focus on a single representation of drugs or proteins. It has been shown that integrating multi-view representations of drugs and proteins can strengthen the prediction ability. For example, a drug can be represented by its chemical structure, or by its chemical response in different cells. A protein can be represented by its sequence, or by its gene expression values in different cells. The docking of drugs and proteins based on their structure can be considered as one view (structural view), and the chemical performance of them based on gene expression and drug response can be considered as another view (chemical view). In this work, we first propose a single-view approach of SLRE based on low rank embedding for an arbitrary view, and then extend it to a multi-view approach of MLRE, which could integrate both views. Our experiments show that our methods perform significantly better than baseline methods including single-view methods and multi-view methods. We finally report predicted drug target interactions for 30 FDA-approved drugs.

Index Terms—Drug target prediction; Multi-view learning; Low rank embedding.

1 INTRODUCTION

In drug discovery process, it is a crucial step to identify drug-target interactions (DTIs). Although the human genome comprises approximately 30,000 genes, proteins encoded by fewer than 400 are used as drug targets in the treatment of diseases. Therefore, it is extremely valuable to discover novel drug targets as the source for drug development. However, due to the high expenses of experiments for validating drug-target interactions, only a small amount of DTIs have been validated by biological experiments. Thus the interest in predicting potential drug-target interactions computationally has recently grown.

Significant work has been done to predict drug-target interactions by using different types of data individually or together, including gene expression data, protein sequence, protein-protein interactions, mass spectrum, chemical structure of drugs, drug response, metabolic network, drug side effects and so on. For example, [1], [2] proposed bipartite graph based methods using protein sequences and drug structures, by first constructing a bipartite graph of drugs and targets, then mapping them to a common space. The close drugs and targets in the space are assumed to be associated with each other. Liu et al.[3] utilize the same data set to predict novel drug-target interactions by their proposed neighborhood regularized logistic matrix factorization, which focuses on modeling the probability that a drug would interact with a target by logistic matrix factorization. Campillos et al. [4] make use of drug side effects to measure the similarity among drugs, which pave a new way to represent drugs. 20 of unexpected DTIs are discovered by their approach, and 13 of them are validated by in vitro binding assays. Mizutani et al.[5] integrated drug side effects and protein function to predict drug targets. Kutalik et al.[6] proposed a modular Ping Pong approach by integrating gene expression and drug response data in NCI-60 cell lines. Drug-gene associations are discovered by identifying co-modules of drugs and genes, which exhibit similar patterns in some cell lines. Drugs and genes in a co-module are assumed to be associated with each other with a high probability. Chen et al. [7] proposed a sparse network-regularized partial least square (SNPLS) method to identify joint modular patterns using large-scale pairwise gene-expression and drug response data, which incorporated a molecular network to the (sparse) partial least square model to improve the module accuracy via a network-based penalty. Ding et al.[8] and Zheng et al. [9] proposed similarity-based methods for predicting drug-target interactions. Li et al. [10] have utilized the human metabolic network as a basis for the prediction of novel targets for known anticancer drugs. Emig et al. [11] proposed a network based approach by integrating disease gene expression signatures and a molecular interaction network. Yuan et al. [12] proposed a novel method, DrugE-Rank, to improve the prediction accuracy by effectively integrating the advantages between feature-based and similarity-based methods which are two types of approaches for drug target prediction. DrugE-Rank uses 'Learning To Rank', for which multiple well-known similarity-based methods can be used as components of ensemble learning.

However, most of the existing methods only consider one type of representation for drugs or proteins. It is still challenging to integrate multi-view data sets for identifying new drug-target associations, since each drug or protein can have multiple representations. For example, a drug can be represented by its chemical structure, or by its chemical response in different cells. A protein can be represented by its sequence, or by its gene expression values in different cells. The docking of drugs and proteins based on their structure can be considered as one view (structural view),

- *Limin Li and Menglan Cai are with School of Mathematics and Statistics, Xi'an Jiaotong University Xi'an, 710049, China. Email: liminli@mail.xjtu.edu.cn,caimenglan@stu.xjtu.edu.cn*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2706267, IEEE/ACM Transactions on Computational Biology and Bioinformatics

and the chemical performance of them based on gene expression and drug response can be considered as another view (chemical view). It could strengthen the prediction ability if we can incorporate the two views. Li [13] proposed a novel single-view graph regularized approach of SPGraph for predicting drug-target interactions by using the structural view data or the chemical view data individually, and then further propose a multi-view co-regularized method by integrating both the structural and chemical representations of drugs and proteins. The results show that the multi-view approach could achieve significantly higher accuracy than single-view approaches. However, the graph Laplacian based methods rely on the graph construction algorithms, and thus are prone to creating false connections among drugs or proteins. In this work, we make use of the same multi-view data sets and propose a single-view approach of SLRE and a multi-view approach of MLRE based on low rank embedding, which is able to avoid the above problem. The experimental results show that our approaches perform significantly better than baseline methods.

The remainder of this paper is structured as follows. In section (3), we first describe the materials used in this paper and introduce the idea of low rank embedding (LRE), then we propose a single view method of SLRE and a multi-view method of MLRE for drug target prediction. In section (4), we first evaluate our two approaches on part of the data sets by comparing them with baseline methods, and report the prediction accuracy values for them, and then apply the multi-view method of MLRE to the whole data sets to predict drug-target interactions.

2 Related work

In this section, we review related work on multi-view learning. Although few work is designed specially for drug target prediction [13], a number of different approaches for multi-view learning have been proposed.

Multi-view approaches can be roughly divided in three families, depending on whether they linearly combine the multiple loss functions in different views[14], [15], [16], or combine multiple graphs with an optimal combination of weights for different purposes [17], [18], [19], [20], or conform different view-specific low-dimensional projections by maximizing the correlations or minimizing the differences [21], [22], [23], [24].

For example, Tang et al.[21] proposed a method of Linked Matrix Factorization(LMF) as a novel way of fusing information from multiple graphs for clustering. In LMF, each graph is factorized as a graph-specific factor and a factor common to all graphs. Wang et al. [24] proposed a similarity network fusion approach for aggregating data types on a genomic scale. Dong et al.[25] proposed two novel methods of SC-GED and SC-SR by efficiently combining the spectrum of the multiple graph layers based on a joint matrix factorization and a graph regularization framework, respectively. In either case, the resulting combination called a joint spectrum of multiple layers was used for clustering. Chen et al. [26] proposed a block spectral clustering method for integrating multiple graphs by constructing block Laplacian matrices. Liu et al. [27] proposed a tensor-based multi-view spectral clustering method with two formulations of SMC-FR-OI and SMC-FR-MI, by solving for shared eigenvectors of tensors based on high-order singular value decomposition and using the computed eigenvectors to cluster.

All the above approaches could not be directly used for the problem of multi-view drug target prediction, where multiple representations for drugs and targets are available, due to the presence of the bipartite graph of drugs and targets. Li [13] proposed a multi-view approach specially for this problem by constructing single-view laplacian eigenmaps first and then apply multi-view co-regularized spectral clustering approaches. However, the construction of the Lapalcian eigenmaps rely on the construction of nearest neighbor graph, and are thus prone to creating spurious inter-manifold connections when the underlying manifolds are mixed. In contrast, low rank embedding is able to avoid this problem and could align data sets drawn from a mixture of manifolds. We propose a multi-view low rank embedding approach for drug target prediction.

3 Materials and methods

In this section, we first explain the multi-view data sets in section (3.1), including the known drug-target interactions, the structural view and chemical view of drugs and proteins. In section (3.2) we describe the problem we want to solve and the notations used in this paper. In section (3.3), we introduce the idea of low rank embedding, which is proposed in [28]. In section (3.4), we then propose a single-view drug target prediction method for an arbitrary view based on the idea of low rank embedding. In section (3.5) we finally propose our multi-view low rank embedding method MLRE for drug target prediction.

3.1 Materials

3.1.1 Drug-target interaction data

We download the known drug-target interactions from the Drug Bank Database[29].

3.1.2 Drug structure data and protein sequence data

We obtain drug structures and protein sequences from KEGG database [30]. We compute the drug structure similarities by using the software of SIMCOMP [31], which is a global alignment algorithm by finding the common substructures between two compounds. We compute the protein sequence similarities using Smith-Waterman method [32] by Matlab.

3.1.3 The NCI60 drug response data and gene expression data

The NCI60 human tumor cell line screen method is developed by National Cancer Institute(NCI) and serves to screen a large number of substances for cytotoxic activity in 60 cell lines for different cancer types. The 60 cell lines were assayed for their sensitivity to a variety of drugs as a part of the Developmental Therapeutics Program(DTP) at the NCI. Each cell line was exposed to each drug for 48 h, and growth inhibition was assessed by the sulforhodamine B assay for cellular protein. The concentration of compound required for 50% growth inhibition was scored as the GI50. We obtain the DTP human tumor cell line screening data from the DTP website. We also obtain gene expression data (mRNA:Affy-U133B, GCRMA-normalized) in NCI-60 cell lines conducted by Shankavaram et al. [33] from NCI website[34].

We construct drug similarities based on drug response data by using a Gauss kernel with σ being the DTP human tumor cell line screening data from the DTP website median distance among all pairwise distances among drugs. Similarly, we use Gauss function to construct gene similarities based on gene expression data.

3.1.4 Data preprocessing and data selection

We first choose 326 overlapping drugs from the drug response data and drug structure data by constructing drug ID mapping between Drug Bank IDs and KEGG IDs. We also obtain 608 overlapping proteins from the gene expression data and protein sequence data. There are totally 114 known interactions among these drugs and protein targets in Drug Bank database. Our final dataset includes 326 drugs and 608 proteins with their known 114 interactions.

Our dataset includes two types of representations for drugs and proteins. The first includes the drug structural similarities and protein sequence similarities data. To interact with a protein target, a drug should have structural docking sites on the protein structure. We thus call this part of the data set the structural view. The other representations are the drug similarities among the 326 drugs based on their response in NCI 60 cell lines, and protein similarities among the 628 proteins based on the expression of their encoding genes in NCI 60 cell lines. Two drugs which show similar profiling in NCI 60 cell lines may interact with the same target, and two genes which show similar expression in 60 cell lines may interact with the same drug. We call this part of the data the chemical view, since the data sets reflect the chemical properties of drugs and proteins. We will use either view and also both views to predict the drug-target interactions.

3.2 Problem Statement and notations

Suppose we have n_d drugs and n_t proteins. We are given the known drug-target interactions, which define a correspondence matrix $W \in \mathcal{R}^{n_d \times n_t}$ as follows

$$W_{ij} = \begin{cases} 1 & \text{drug } i \text{ is known to have protein target } j, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose the similarities among drugs or proteins in the structural view are denoted by $K_d^{(s)}$ or $K_t^{(s)}$, and the similarities among drugs or proteins in the chemical view are denoted by $K_d^{(c)}$ or $K_t^{(c)}$[1]. A typical drug target prediction problem aims to find new drug-target interactions based on the above information.

We first focus on an arbitrary view and use $K_d \in \mathcal{R}^{n_d \times n_d}$ and $K_t \in \mathcal{R}^{n_t \times n_t}$ to represent the similarity matrices among drugs and proteins, respectively. We then aim to predict drug-target interactions by integrating both the structural and chemical views.

[1] 's' and 'c' are the short for structural and chemical,respectively.

3.3 Low rank embedding

Low rank embedding (LRE) is proposed in [28] to project n samples $X = [x_1, \cdots, x_n] \in \mathcal{R}^{p \times n}$ in a lower-dimensional subspace as $Z = [z_1, \cdots, z_n] \in \mathcal{R}^{k \times n}$, where $k < p$, such that the local structure of the original samples are preserved. Low rank embedding is a two-step algorithm. The first step is to find a low-rank representation R of the data X by the following optimization problem

$$\min_R \|X - XR\|_F^2 + \lambda_0 \|R\|_* \tag{1}$$

where $\|R\|_* = \sum_i \sigma_i(R)$ is the nuclear norm, for singular values σ_i, and λ_0 is a regularization parameter controlling the reconstruction error in the first term and the low rank property in the second term. λ_0 is usually chosen as 1. The optimization problem can be solved by the alternating direction method of multipliers (ADMM) [35] or by computing the singular value decomposition of X[36]. Suppose the singular value decomposition of X is $X = USV^T$, and V and S are partitioned into $V = [V_1 \ V_2]$ and $S = \left(\begin{array}{cc} S1 & 0 \\ 0 & S2 \end{array} \right)$ such that S_1 includes the singular values larger than 1, and S_2 includes the ones smaller than or equal to 1. Then the solution R of problem (1) can be computed by $R = V_1(I - S_1^{-2})V_1^T$.

The second step of LRE is to fix R and minimize the reconstruction error in the embedded space such that the point-wise linear reconstruction is preserved,

$$\min_Z \|Z - ZR\|_F^2, s.t.\ ZZ^T = I \tag{2}$$

where Z is the embedding of X, and the constraint $ZZ^T = I$ ensures that the problem is well-posed. The problem can be rewritten as

$$\min_{ZZ^T=I} tr(Z(I-R)(I-R)^T Z^T) \tag{3}$$

and thus the solution Z can be obtained by the eigenvectors of $(I-R)(I-R)^T$ corresponding to the k smallest eigenvalues.

To predict drug target prediction, we hope that drugs and targets are lying in the same distance space such that the distance among drugs and targets can be used to measure the strength of their interactions. Therefore, we need embed both drugs and targets in a common low-dimensional subspace with some constraints. Low rank embedding is a relatively new and effective method for dimension reduction. Thus we adapted LRE to embed drugs and targets simultaneously with a novel constraint based on known drug-target interactions. Furthermore, Li [13] proposed a multi-view approach specially for this problem by constructing single-view laplacian eigenmaps first and then apply multi-view co-regularized spectral clustering approaches. However, the Lapalcian eigenmaps rely on the construction of nearest neighbor graph, and are thus prone to creating spurious inter-manifold connections when the underlying manifolds are mixed. In contrast, low rank embedding is able to avoid this problem and could align data sets drawn from a mixture of manifolds. This also motivates us to use LRE for the problem of drug target prediction.

The difficulty to directly use low rank embedding for drug target prediction is two-fold. First, in our problem, the

Algorithm SLRE of DTI prediction:

```
Inputs. $K_d$,$K_t$, $W$
        $\lambda, \mu, k$
Outputs. $\tilde{Z}$
1. Compute the eigenvalue decomposition of $K_t$ and $K_d$.
2. Compute $R_t$ and $R_d$ from (5)
3. Construct $\tilde{R}$ and $\tilde{W}$ from (7)
4. Compute $M$ from (9)
5. Compute the eigenvalue decomposition of $M$, and the rows of $Z$
   are set to be the eigenvectors of $M$ corresponding to its $k$ smallest
   eigenvalues
6. Apply $k$-means on the columns of $\tilde{Z}$ to do clustering for drugs
   and proteins.
```

Algorithm MLRE of DTI prediction:

```
Inputs. structural view: $K_d^{(s)}$,$K_t^{(s)}$,
        chemical view: $K_d^{(c)}$,$K_t^{(c)}$
        known drug-target interactions: $W$,
        $\lambda, \mu^{(s)}, \mu^{(c)}, k$
Outputs. $Z^{(s)}, Z^{(c)}$
1. Compute $M^{(s)}$ using line (1-4) in algorithm SLRE, with $K_d^{(s)}$,$K_t^{(s)}$
   and $\mu^{(s)}$ in the structural view.
2. Compute $M^{(c)}$ using line (1-4) in algorithm SLRE, with $K_d^{(c)}$,$K_t^{(c)}$
   and $\mu^{(c)}$ in the chemical view.
3. Solve model (10) for $Z^{(s)}$ and $Z^{(c)}$ by alternating optimization.
4. Apply $k$-means on the columns of $Z^{(s)}$ to do clustering for drugs
   and proteins.
```

vector representation for each drug or target x_i for structural view is unknown, and only the similarities among drugs or proteins are given. Second, the known interactions between drugs and proteins should be integrated in the model. We will propose our methods in the following subsection.

3.4 drug target prediction by low rank embedding

In this section, we focus on an arbitrary view and use $K_d \in R^{n_d \times n_d}$ and $K_t \in R^{n_t \times n_t}$ to represent the similarity matrices among drugs and proteins for the view, respectively. We propose a single-view method for drug target prediction based on low rank embedding. Our main idea is to first compute the reconstruction matrices R_d and R_t for drugs and proteins, respectively, and then embed them in a common low-dimensional space such that their pointwise linear reconstructions are preserved and the known drug-target interactions are best preserved. We propose a single-view method SLRE for drug target prediction, which includes two steps as follows.

In the first step, we compute reconstruction matrices R_d and R_t for drugs and proteins, respectively. Note that the vector representations for drugs or targets are unavailable for the structural view and only the similarities among drugs or proteins are given, thus the low rank embedding model in (1) can not be used directly. We propose a kernel version low rank embedding as follows.

$$\min_{R} tr((I-R)^T K (I-R)) + \lambda \|R\|_* \tag{4}$$

where K can be chosen as K_d or K_t, R can be R_d or R_t for drugs and proteins, respectively. To solve the problem, we only need to compute the eigenvalue decomposition of $K = VS^2V^T$, partition V and S in the same way as the above subsection, and finally compute

$$R = V_1(I - S_1^{-2})V_1^T. \tag{5}$$

In the second step, we fix R_d and R_t computed in the first step, and project drugs and targets into a common k-dimensional space such that their linear reconstructions are preserved, and the known drug-target interactions are best captured. The model is proposed as

$$\begin{aligned} \min_{Z_d, Z_t} & \|Z_d - Z_d R_d\|_F^2 + \|Z_t - Z_t R_t\|_F^2 \\ & + \mu \sum_{i,j} \|z_{d,i} - z_{t,j}\|_2^2 W_{ij} \\ & s.t. Z_d Z_d^T = Z_t Z_t^T = I_k \end{aligned} \tag{6}$$

where $z_{d,i}$ represents the i-th column of Z_d and the k-dimensional representation of the drug i, and $z_{t,j}$ represents the k-dimensional representation of the protein j, μ controls the trade-off between the low-rank embedding and the known drug-target interactions recovery in the low-dimensional common space.

To solve the problem in (6), we define the block matrices $\tilde{R}, \tilde{W} \in R^{n \times n}$, $\tilde{Z} \in R^{k \times n}$ as

$$\tilde{R} = \begin{bmatrix} R_d & 0 \\ 0 & R_t \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} 0 & W \\ W^T & 0 \end{bmatrix}, \quad \tilde{Z} = \begin{bmatrix} Z_d & Z_t \end{bmatrix}, \tag{7}$$

where $n = n_d + n_t$ and k is the dimension of the common space. Then the model in equation (6) can be changed into

$$\begin{aligned} \min_{\tilde{Z}} & \|\tilde{Z} - \tilde{Z}\tilde{R}\|_F^2 + \mu \textstyle\sum_{i,j} \|\tilde{z}_i - \tilde{z}_j\|_2^2 \tilde{W}_{ij} \\ & s.t. \tilde{Z}\tilde{Z}^T = I \end{aligned} \tag{8}$$

where $\tilde{z}_i$ and $\tilde{z}_j$ is the i-th and j-th column in Z respectively. Similarly, the constrain is used to ensure the problem to be well-posed.

The objective in equation (8) can be further rewritten as

$$\begin{aligned} & \|\tilde{Z} - \tilde{Z}\tilde{R}\|_F^2 + \mu \textstyle\sum_{i,j} \|\tilde{z}_i - \tilde{z}_j\|_2^2 W_{ij} \\ = \; & tr(\tilde{Z}(I - \tilde{R})(I - \tilde{R})^T \tilde{Z}^T) + 2\mu tr(\tilde{Z} L \tilde{Z}^T) \\ = \; & tr(\tilde{Z}((I - \tilde{R})(I - \tilde{R})^T + 2\mu L)\tilde{Z}^T), \end{aligned}$$

where $L = \tilde{D} - \tilde{W}$, $\tilde{D}$ is a diagonal matrix with $\tilde{D}_{ii} = \sum_j \tilde{W}_{ij}$. We denote

$$M = (I - \tilde{R})(I - \tilde{R})^T + 2\mu L, \tag{9}$$

and the optimal solution of problem in (8) can be obtained by the eigenvectors of M corresponding to its k smallest eigenvalues. We finally apply k-means on the new representations $\tilde{Z}$ of drugs and proteins in the common space, and predict drugs and proteins in the same cluster to be interacted with each other. Similar to spectral clustering, the number of clusters is chosen as the low dimensionality k. The detail of the algorithm is shown in the algorithm box of SLRE.

3.5 Multi-view low rank embedding for drug target prediction

In this section, we further extend SLRE to a multi-view method MLRE which integrates both the structural and chemical views of drugs or proteins. In the structural view, a drug is represented by its chemical structure, and a protein is represented by its amino-acid sequence. In the chemical view, a drug is represented by its chemical response in different cells, and a protein is represented by its gene

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2017.2706267, IEEE/ACM Transactions on Computational Biology and Bioinformatics

expression levels in different cells. We could predict drug-target interactions using either view individually. However, the integration of the two views may improve the prediction accuracy.

Suppose in the structural view of drugs and proteins, we construct $M^{(s)}$ in equation (9) with the single-view low rank embedding method in the above subsection. We construct $M^{(c)}$ for the chemical view in the same way. Suppose $Z^{(s)}$ and $Z^{(c)}$ are the embedding of the drugs and targets in k-dimensional space for the structural view and chemical view, respectively. Borrowing the idea of co-regularized spectral clustering [37], we hope the embedding representations $Z^{(s)}$ and $Z^{(c)}$ to be as close as possible. Similar to the MPGraph approach in [13], our multi-view low-rank embedding model is proposed as follows

$$\begin{aligned} \min_{Z^{(s)},Z^{(c)}} \; & tr(Z^{(s)}M^{(s)}{Z^{(s)}}^T) + tr(Z^{(c)}M^{(c)}{Z^{(c)}}^T) \\ & -\lambda tr({Z^{(s)}}^T Z^{(s)} {Z^{(c)}}^T Z^{(c)}) \\ s.t. \;\; & Z^{(s)}{Z^{(s)}}^T = I, Z^{(c)}{Z^{(c)}}^T = I, \end{aligned} \tag{10}$$

where λ controls the trade-off between the goodness of embedding in each single view and the consistency between the two embedding. The last term in the objective function of (10) is called Hilbert-Schmidt Independence Criterion (HSIC), which could measure the similarity between two representations. The optimization problem in equation (10) can be solved by alternating optimization. For a given $Z^{(s)}$, we get the optimization

$$\min_{Z^{(c)}{Z^{(c)}}^T=I} tr(Z^{(c)}(M^{(c)} - \lambda {Z^{(s)}}^T Z^{(s)}){Z^{(c)}}^T). \tag{11}$$

The solution $Z^{(c)}$ can be obtained by the top eigenvectors of $M^{(c)} - \lambda {Z^{(s)}}^T Z^{(s)}$ corresponding to the k smallest eigenvalues. Once $Z^{(c)}$ is obtained, we can update $Z^{(s)}$ in the same way. The convergence stops when the differences in the values of objective function between consecutive iterations fall below a minimum threshold $\epsilon \leq 1e-4$. Either $Z^{(s)}$ or $Z^{(c)}$ can be used for k-means to do clustering for drugs and targets. The number of clusters is chosen as the low dimensionality k. The detail of the algorithm is shown in the algorithm box of MLRE.

3.6 Complexity analysis

At each iteration of our algorithm MLRE, Z is updated in two steps for each view. At the first step, singular value decomposition on data K_d or K_t is computed to construct the matrices R_d or R_t and M, and a computation cost of O(n_d^3) or O(n_t^3) is required. At the second step, Z is updated by the eigenvectors of M corresponding to its k smallest eigenvalues, which requires a computation cost of O(n_d^3) or O(n_t^3). Overall, this algorithm takes computation time of O(n_d^3) or O(n_t^3).

4 Experimental Results

4.1 Evaluation of the MLRE approach

We did experiments to show the performance of our approaches, including SLRE-S, SLRE-C, and MLRE. SLRE-S and SLRE-C represents the single-view SLRE approach applied to the structural view and the chemical view, respectively. To evaluate our methods, we compare with several baseline methods and compute the prediction accuracy for all these methods. We first describe these baseline methods, then explain the experimental setting, and finally show the experimental results.

4.1.1 Baseline methods

a). *Single-view and multi-view SVMs*. Support vector machines(SVMs) can be used to classify drug-target pairs to 'interacted' or 'not interacted', based on training data sets. We use the Kronecker product $K = K_d \otimes K_t$ of the drug kernel K_d and protein kernel K_t as the kernel between drug-protein pairs. We first use SVM with these Kronecker kernels for the structural view and the chemical view individually, and then apply SVM with multiple kernel for both views.

b). *Ping Pong algorithm (PP)[6]*. Ping Pong algorithm is a modular approach which could integrate gene expression data and drug response data for drug target prediction. Thus it is a single-view approach only for the chemical view. Note that PP is an unsupervised method since the known drug-target interactions are not used.

c). *Bipartite graph learning(BGL)[1]*. Bipartite graph learning is a single-view method which uses structural view or chemical view for drug target prediction.

d). *SPGraph and MPGraph [13]*. SPGraph is a single-view method which could use either structural view (SPGraph-S) or chemical view (SPGraph-C) for drug target prediction. MPGraph is a multi-view approach which integrates both views for drug target prediction.

4.1.2 Experimental setting

To evaluate our method, we choose a smaller subset of our data, in which each drug has at least one known protein target and each protein is known to be a target of at least one drug, and totally there are 65 drugs and 80 targets with the 114 interactions. On this smaller subset of data, we evaluate our methods in three experimental settings, new drug(ND), new target(NT), and new drug-new target(NDNT). In the setting of ND, we aim to find out the new interactions between test drugs and all proteins. In the setting of NT, we want to find out the new interactions between test proteins and all drugs. In the setting of NDNT, we want to find out the new interactions between test drugs and test proteins. For ND, the drug data are randomly partitioned into 5-folds. The drugs in each fold is considered as test drugs once while ones in the remaining folds are considered as training drugs. Using the known interactions between the training drugs and all proteins, we can embed all drugs and proteins to a k-dimensional space by SLRE-S, SLRE-C, or MLRE, where the closest t proteins with regard to linear kernel $K = Z^T Z$ around a test drug are possible to be its targets. By changing the threshold t, one can get the ROC curve and compute the AUC values. Z is chosen as $Z^{(s)}$ or $Z^{(c)}$ in our experiments and the best result is reported. The setting of NT is similar to ND, and the only difference is that we partition all the proteins not drugs into training and test sets. In the setting

TABLE 1
The average AUCs by the SLRE, MLRE and other comparison partners.

	Structure				Chemical					Multi-view			
AUCs	SVM	BGL	SPGraph	SLRA	PP	SVM	BGL	SPGraph	SLRA	MKL-SVM	MPGraph	MLRE	CPUtime(s)
ND	0.576 ± 0.003	0.553 ± 0.003	0.560 ± 0.011	**0.601 ± 0.009**	0.478 ± 0.010	0.471 ± 0.002	0.505 ± 0.003	**0.525 ± 0.017**	0.503 ± 0.005	0.518 ± 0.003	0.627 ± 0.009	**0.650 ± 0.046**	0.0582
NT	0.502 ± 0.002	0.443 ± 0.003	**0.509 ± 0.006**	0.499 ± 0.006		0.486 ± 0.002	0.499 ± 0.003	**0.541 ± 0.009**	0.507 ± 0.005	0.522 ± 0.003	**0.552 ± 0.006**	0.514 ± 0.012	0.0597
NDNT	0.508 ± 0.009	0.480 ± 0.011	0.527 ± 0.027	**0.590 ± 0.011**		0.467 ± 0.010	0.497 ± 0.009	0.478 ± 0.008	0.431 ± 0.011	0.502 ± 0.010	0.581 ± 0.019	**0.617 ± 0.060**	0.0577

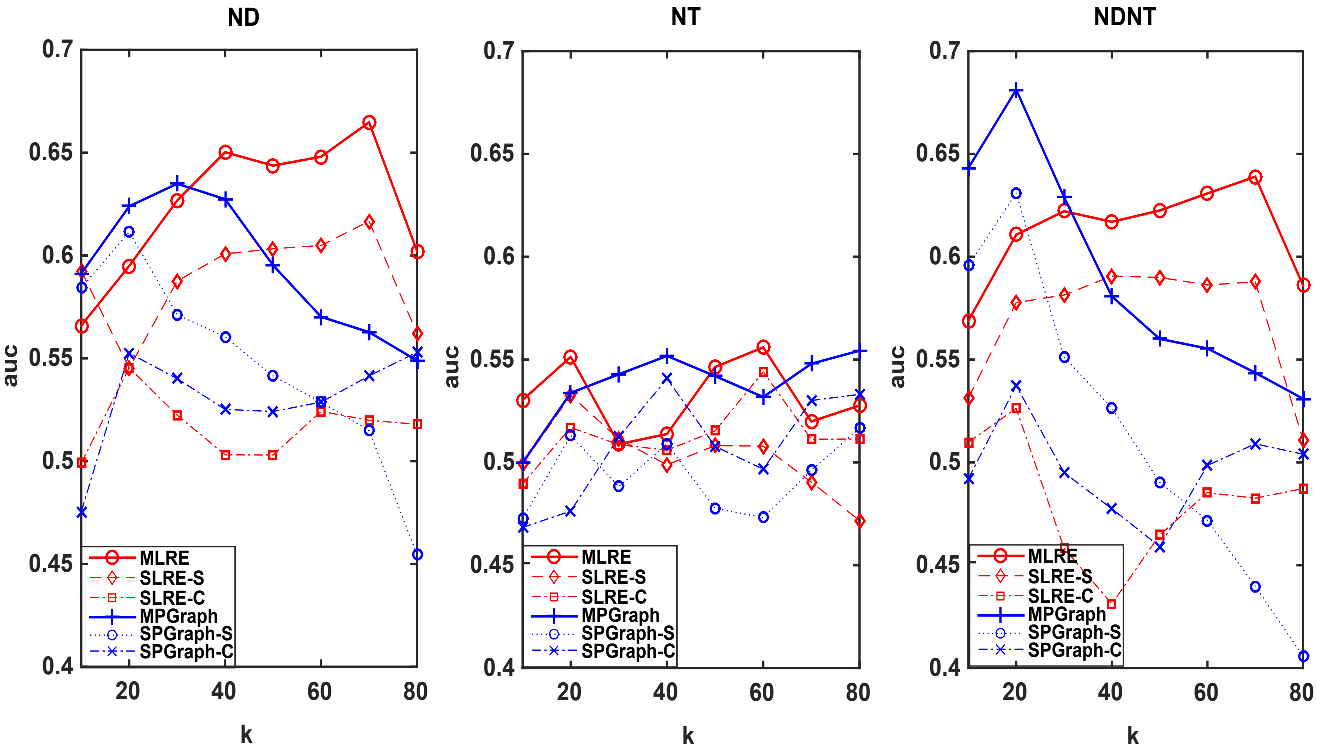

Fig. 1. The average AUC results computed by six approaches in three settings of ND, NT and NDNT.

of NDNT, we partition both drugs and proteins into 5-folds, respectively. Each time we choose 1-fold of drugs and 1-fold of proteins as the test data, and the remaining as the training data. We embed all drugs and proteins into a k-dimensional space by using the known interactions between the training drugs and the training proteins. We use the same way as the case of ND to calculate the AUC values. In each setting, we randomly partitioned data 50 times, and report the average AUCs and the standard errors. Note that the final step of k-means in algorithms SLRE and MLRE is not used in these experiments.

For both SLRE and MLRE methods, we found that the parameters of μs are robust in a relatively large range, and thus we set $\mu^{(s)} = \mu^{(c)} = 0.1$. For MLRE method, we report the best results when λ is chosen from $\{0.01, 1, 100\}$. The dimension k is fixed as 40 first and then we show its robustness by reporting the results with different ks.We use the same parameter setting of $\mu^{(s)}$, $\mu^{(c)}$, λ and k for SPGraph and MPGraph approaches.

For Ping-Pong method, we compute the co-modules for all the 65 drugs and 80 targets based on the data in the chemical view: drug response and gene expression in the NCI60 cell lines. New drug-target associations are discovered if they belong to the same co-module. Since Ping Pong algorithm depends on random initials, We repeat the computation 50 times, and report the average AUC and standard error. Note that this method can not be used for the structural data, so we can only implement it on the chemical view.

4.1.3 Results

We first report the results for our approaches with $k = 40$ and baseline methods in Table 1. We can see that single-view approaches with the structural view generally obtained higher AUCs than those with the chemical view, except that in NT setting, all the approaches performed very bad, either in chemical view or in structural view. For single-view approaches with the structural view, our method of SLRA performed best for the ND and NDNT settings. For these approaches with the chemical view, all methods performed bad, and SPGraph worked the best. Note that, For MLRE, if we use $Z^{(s)}$ for each cross validation, the average AUCs are 0.624, 0.506 and 0.582 for the settings of NT, ND, and NDNT, respectively. If we use $Z^{(c)}$, the average AUCs are 0.629, 0.506 and 0.574 for the settings of NT, ND, and NDNT, respectively. In Table 1, we reported the best result by using $Z^{(s)}$ or $Z^{(c)}$ for each cross validation, for both MPGraph and MLRE. We can see from the table that, In the settings of ND and NDNT, graph based multi-view method (MPGraph) worked better than the corresponding single-view method (SPGraph), and our proposed low rank embedding based multi-view method (MLRE) also worked better than the corresponding single-view method (SLRE). This implies that using multi-view representations of drugs and proteins

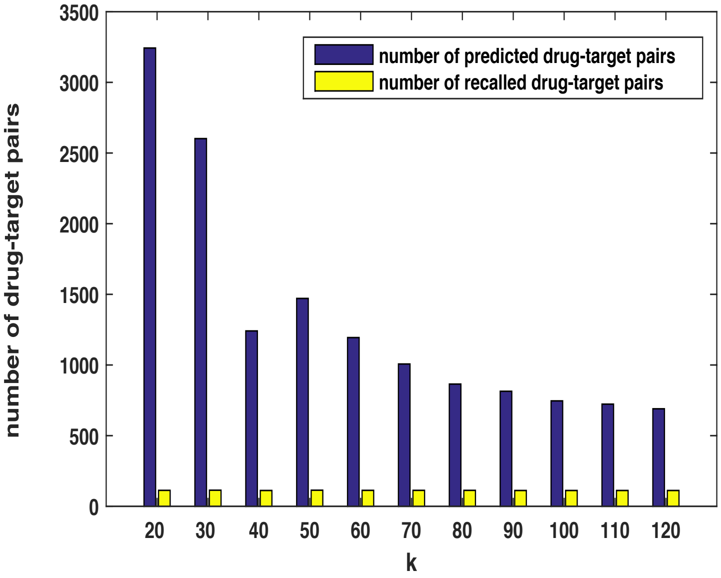

Fig. 2. The number of drug-target pairs for different ks.

could strengthen the prediction ability. We also note that for multi-view methods, our proposed MLRE worked the best for the settings of ND and NDNT, and MPGraph worked the best for the NT setting. This shows that although the SLRE didn't improve the accuracy in the chemical view, but its multi-view version MLRE could significantly improve the prediction accuracy.

To see the robustness of our approaches for the parameter of k, we fixed k from the set of $\{10 : 10 : 80\}$ and showed the results of SPGraph-S, SPGraph-C, MPGraph, SLRE-S, SLRE-C,and MLRE for the three settings in Figure 1. From the figure, we can see that, for any k from the parameter set, single-view approaches with the structural view generally work better than those with the chemical view, and the multi-view methods generally work better than the single-view methods. We also see that, our SLRE and MLRE approaches are less sensitive than the SPGraph and MPGraph. For example, in the NDNT setting, the accuracy obtained by MPGraph drops sharply when k increases, but our MLRE performs very stably when k is in a relatively large set. From Figure 1, we can also see that, both MLRE and MPGraph obtained the best and stable results choosing k as 40, and thus k is set to be 40 in Table 1.

4.2 drug target prediction using the whole data set

We applied our multi-view approach MLRE on both the structural view and the chemical view of the original data with $n_d = 326$ drugs and $n_t = 608$ proteins to predict novel drug-target interactions. The parameters are set as $\mu^{(s)} = \mu^{(c)} = 1$, $\lambda = 0.01$ and $k = 40$ by the experience in the above experiments. We applied the MLRE algorithm to project drugs and proteins to a k-dimensional Euclidean space and then cluster them to co-modules. Note that the result of k-means depends on the intials. To obtain a stable result, we conducted 100 times of k-means with random initials and thus obtained 100 different clustering results. Each clustering result corresponds to a membership matrix $G_t \in R^{n_d \times n_t}, t = 1, \cdots, 100$, where $G_t(i,j) = 1$ means drug i and protein j are in the same cluster, and $G_t(i,j) = 0$ otherwise. The average membership matrix is defined as $G = \sum_{t=1}^{100} G_t/100$, which corresponds to a bipartite graph of drugs and proteins. We further remove the edges with weights less than 0.6 in G and the new graph is denoted as $\mathcal{G}$. Thus the drug-target pairs linked in $\mathcal{G}$ belong to one co-module with enough confidence, and can be considered as drug-target interactions. Note that most of the known drug-target interactions are recalled among the edges in $\mathcal{G}$, which is shown in Figure 2. we reported the number of edges in $\mathcal{G}$ and the number of the recalled interactions for $k = 20 : 10 : 120$. We can see that the majority of the known 114 interactions can be recalled.

TABLE 2
The predicted targets for 30 FDA-approved drugs.

KEGG ID	Drug Name	Gene Name
D00341	Hydroxyurea	NSDHL,MAT2A,STAT1
D00363	Lomustine	FARSLA,CLPP
D00211	Rifampin	PIK3R1,VCAM1,UCK2,NR1H2,JARID1D
D01211	Leucovorin	GRM3
D00254	Carmustine	FARSLA
D05932	Streptozocin	FARSLA,CLPP
D00609	Prazosin	SDS,KIT,MAPK3,SDHC
D00433	Silver sulfadiazine	POLA2,CTNNB1
D05096	Mycophenolic acid	PDE2A
D01324	Mequitazine	KIF1A
D04066	Estramustine	PIK3R1,JARID1D
D00094	Tretinoin	GABRR1,CAT,CTBP1
D01352	Dinoprost Tromethamine	NR0B1,APRT,OPRD1,GSTM3, GLO1,MAOA
D00214	Dactinomycin	PIK3R1,VCAM1,UCK2,NR1H2,JARID1D, CRYBB1
D02321	Amsacrine	CP
D00496	Penicillamine	HADHSC,SLC1A4
D00473	Prednisone	ALDH3B2
D00467	Pipobroman	FARSLA,CLPP
D04872	Mechlorethamine	FARSLA,CLPP
D02671	Mesoridazine	UROD
D00554	Ethinyl Estradiol	PIK3R1,UCK2,NR1H2,JARID1D,CRYBB1
D01745	Domperidone	UROD
D02335	Menadione	RHO
D00142	Methotrexate	GRM3
D00184	Cyclosporine	PIK3R1,VCAM1,UCK2,NR1H2
D00221	Acetylcysteine	HADHSC,SLC1A4
D00570	Colchicine	VCAM1,UCK2,NR1H2,JARID1D
D00188	Cholecalciferol	PIK3R1,VCAM1,UCK2,NR1H2, JARID1D,CRYBB1
D00364	Loratadine	KIF1A
D00153	Testolactone	ALDH3B2

We finally reported in Table 2 the new predicted drug-target interactions for 30 FDA-approved drugs (i.e. Food and Drug Administration-approved drugs) with weights 1 in graph $\mathcal{G}$ when k is chosen as 40. We observed that some of the predicted drug-target interactions in Table 2 can be supported by some references, and some predicted interactions can give us some hints for understanding the drug mechanism or side effects.

Mycophenolic acid, an antineoplastic drug, is known to be an inhibitor of inosine monophosphate dehydrogenase (IMPDH). We predict PDE2A(Phosphodiesterase 2A, CGMP-Stimulated) to be a possible target for Mycophenolic acid. PDE2A might have some link with Mycophenolic acid, since the gene PDE2A belongs to a family of related phosphohydrolyases that selectively catalyze the hydrolysis of 3′ cyclic phosphate bonds in adenosine and/or guanine cyclic monophosphate (cAMP and/or cGMP). Acetylcysteine is known to have antiviral effects in HIV patients since viral stimulation can be inhabited by reactive oxygen intermediates. Our results show HADHSC and SLC1A4 could be targets of Acetylcysteine. HADHSC (Hydroxyacyl-CoA Dehydrogenase), a member of the 3-hydroxyacyl-CoA dehydrogenase gene family, can catalyze the oxidation of straight-chain 3-hydroxyacyl-CoAs in the mitochondrial matrix as part of the beta-oxidation pathway. The interaction between HADHSC and Acetylcysteine is discussed in [38], [39], and also predicted in [13]. Besides, SLC1A4(Solute Carrier Family 1, Member 4) might be a transporter of Acetylcysteine, since another gene SLCO1B1 in solute carrier family is known to be one of its transporters. Testolactone is an antineoplastic agent for the treatment of breast cancer, and it has validated target Cytochrome P450 19A1, which catalyzes the formation of aromatic C18 estrogens from C19 androgens [40]. Our results show ALDH3B2 (Aldehyde Dehydrogenase 3 Family, Member B2) could be a target of this drug, and we found that ALDH3B2 is involved in the pathway of Drug metabolism - cytochrome P450 [30]. Hydroxyurea is an antineoplastic agent. It inhibits DNA synthesis through the inhibition of ribonucleoside diphosphate reductase. Besides, hydroxyurea has the suppression function on Sterol Biosynthesis [41]. We predicted NSDHL, MAT2A and STAT1 might be its targets. Among the proteins, NSDHL is an enzyme involved in cholesterol biosynthesis that may be inhibited by Hydroxyurea. Mechlorethamine is used as an antineoplastic in Hodgkins disease and lymphomas. The target predicted in the table is FARSLA, which is known to be related with tRNA. We found that in B cell lymphoma tRNA-derived microRNA is down-regulated which modulates proliferation and the DNA damage response [42]. By the tRNA, mechlorethamine and FARSLA may have some interactions. Prostate neoplasms is usually treated by Estramustine which has radiation protective properties. We predicted two targets for Estramustine: PIK3R1 and JARID1D. PIK3R1 is a regulatory subunit gene of the PI3K pathway, and it has been proved that PIK3R1 is under direct control of androgens and can be repressed in prostate cancer cells [43]. When prostate cancer invades or metastasizes, JARID1D can be the inhibition and prognostic marker [44]. Amsacrine is an antineoplastic agent for the treatment of acute leukemias and malignant lymphomas. We predicted CP as its target. It is known that CP is involved in ferroxidase activity. In the article [45], we found that CP-1 mRNA is the predominant CP transcript in a lymphoblastic leukemia cell line.

Based on the above discussion, we showed that our predicted drug-target interactions could be supported by references, and could provide useful hints for further biological study of drug mechanisms.

5 Conclusion

Multi-view approaches have been shown to be effective for many applications when multiple representations of objects can be obtained. In this work, we first develop a single-view approach of SLRE to predict drug-target interactions by integrating drug structures and protein sequences, or integrating drug responses and gene expression levels in NCI 60 cancer cell lines. We further extend SLRE method to a multi-view approach of MLRE, which could integrate both the structural view and chemical view of the drugs and proteins. Experimental results show that our approaches could significantly improve the prediction accuracy in most cases.

Note that although we only consider two views of drugs and proteins, our MLRE can be extended to the case for more than two views. The drug side effect is another type of data source widely used in drug target prediction[4], [46], and biological function in gene ontology[47] of off-targets often result in the side effects of the drugs. Thus three views of drugs and targets can be used for drug target prediction: structural view, chemical view, and functional view. It is an interesting topic to extend MLRE to three views to strengthen the prediction ability. This might be a research topic in the future.

As we can see, the average AUC values reported in Table 1 as a whole are not good enough. The reasons for this might be two-fold. On the one hand, the dataset we used for reporting table 1 contains a limited number of drugs, targets and interactions(only 65 drugs, 80 targets and 114 known drug-target interactions). The five-fold cross validation makes an even smaller training dataset. It is very difficult to obtain a high AUC value with such a small dataset. On the other hand, this could be due to the presence of view disagreement of multiple views, i.e, when samples in each view do not belong to the same class due to view corruption, occlusion, other noise processes. A common assumption in multi-view learning is that the samples from each view always belong to the same class. In realistic settings, datasets are often corrupted by noise. Multi-view learning approaches have difficulty dealing with noisy observations, especially when each view is corrupted by an independent noise process. In our future work, we hope to address the problem when the view disagreement is present.

Another interesting future research topic can be the multi-view learning with missing values. Note that our approach can only be applied for complete multi-view data, and thus a lot of drugs or genes have to be thrown away if the data for either view is missing. This clearly limits the applications of MLRE method. Effective multi-view approaches with missing values will be considered in our future research.

6 Acknowledgment

The work was supported by the NSFC projects 11471256 and 11631012.

References

[1] Yamanishi.Y, Araki.M.A, Honda.W, and Kanehisa.M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24(13):i232–i240(9), 2008.

[2] Bleakley.K and Yamanishi.Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–403, 2009.

[3] Liu.Y, Wu.M, Miao.C, Zhao.P, and Li.X. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *Plos Computational Biology*, 12(2):7762–7773, 2016.

[4] Campillos.M, Kuhn.M, Gavin.A.C, Jensen.L.J, and Bork.P. Drug target identification using side-effect similarity. *Science*, 321(5886):263–6, 2008.

[5] Mizutani.S, Pauwels.E, Stoven.V, Goto.S, and Yamanishi.Y. Relating drugprotein interaction network with drug side effects. *Bioinformatics*, 28(18):i522–i528, 2011.

[6] Kutalik.Z, Beckmann.J.S, and Bergmann.S. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology*, 26(5):531–539, 2008.

[7] Chen.J and Zhang.S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics*, 32(11), 2016.

[8] Ding.H, Takigawa.I, Mamitsuka.H, and Zhu.S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. *Briefings in Bioinformatics*, 15(5):734–747, 2014.

[9] Zheng.X, Ding.H, Mamitsuka.H, and Zhu.S.F. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. 2013.

[10] Li.L, Zhou.X, Ching.W.K, and Wang.P. Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in nci-60 cell lines. *BMC Bioinformatics*, 11(1):501, 2010.

[11] Emig.D, Ivliev.A, Pustovalova.O, Lancashire.L, Bureeva.S, Nikolsky.Y, and Bessarabova.M. Drug target prediction and repositioning using an integrated network-based approach. *Plos One*, 8(4):e60618–e60618, 2013.

[12] Yuan.Q, Gao.J, Wu.D, Zhang.S, Mamitsuka.H, and Zhu.S. Drugerank: improving drugtarget interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, 32(12):i18–i27, 2016.

[13] Li.L. Mpgraph: multi-view penalised graph clustering for predicting drug-target interactions. *Iet Systems Biology*, 8(2):67–73, 2014.

[14] Shen.R, Mo.Q, Schultz.N, Seshan.V.E, Olshen.A.B, Huse.J, Ladanyi.M, and Sander.C. Integrative subtype discovery in glioblastoma using icluster. *Plos One*, 7(4):e35236, 2012.

[15] Mo.Q, Wang.S, Seshan.V.E, Olshen.A.B, Schultz.N, Sander.C, Powers.R.S, Ladanyi.M, and Shen.R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, 110(11):4245, 2013.

[16] Gnen.M and Margolin.A.A. Localized data fusion for kernel k-means clustering with application to cancer biology. *Advances in Neural Information Processing Systems*, 2:1305–1313, 2014.

[17] Lanckriet.G.R.G, Cristianini.N, Bartlett.P, El.G.L, and Jordan.M.I. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5(1):27–72, 2002.

[18] Yu.S, Tranchevent.L.C, Liu.X, and Glanzel.W. Optimized data fusion for kernel k-means clustering. *Pattern Analysis and Machine Intelligence IEEE Transactions on*, 34(5):1031–1039, 2011.

[19] Lange.T and Buhmann.J.M. Fusion of similarity data in clustering. *Inproceeding of Advances in Neural Information Processing Systems*, 2005.

[20] Chuang.Y. Affinity aggregation for spectral clustering. 23(10):773–780, 2012.

[21] Tang.W, Lu.Z, and Dhillon.I.S. Clustering with multiple graphs. 24(4):1016–1021, 2009.

[22] Chaudhuri.K, Kakade.S.M, Livescu.K, and Sridharan.K. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, 2009.

[23] Kumar.A, Rai.P, and Daum.H. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, 2012.

[24] Wang.B, Mezlini.A.M, Demir.F, Fiume.M, Tu.Z, Brudno.M, Haibekains.B, and Goldenberg.A. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333, 2014.

[25] Dong.X, Frossard.P, Vandergheynst.P, and Nefedov.N. Clustering with multi-layer graphs: A spectral perspective. *IEEE Transactions on Signal Processing*, 60(11):5820–5831, 2011.

[26] Chen.C, Ng.M.K, and Zhang.S. Block spectral clustering methods for multiple graphs. *Numerical Linear Algebra with Applications*, 2016.

[27] Liu.X, Ji.S, Glanzel.W, and De.M.B. Multiview partitioning via tensor methods. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):1056–1069, 2013.

[28] Liu.R, Hao.R, and Su.Z. Mixture of manifolds clustering via low rank embedding. *Journal of Information and Computational Science*, 8(5):725–737, 2011.

[29] Knox.C, Law.V, Jewison.T, Liu.P, Ly.S, Frolkis.A, Pon.A, Banco.K, Mak.C, and Neveu.V. Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39(suppl1):1035–41, 2011.

[30] Kanehisa.M, Goto.S, Hattori.M, Aokikinoshita.K.F, Itoh.M, Kawashima.S, Katayama.T, Araki.M, and Hirakawa.M. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Research*, 34(suppl1):354–357, 2006.

[31] Hattori.M, Okuno.Y, Goto.S, and Kanehisa.M. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–65, 2003.

[32] Smith.T.F and Waterman.M.S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195197, 1981.

[33] Shankavaram.U.T, Reinhold.W.C, Nishizuka.S, Major.S, Morita.D, Chary.K.K, Reimers.M.A, Scherf.U, Kahn.A, and Dolginow.D. Transcript and protein expression profiles of the nci-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3):820–832, 2007.

[34] Reinhold.W.C, Sunshine.M, Liu.H, Varma.S, Kohn.K.W, Morris.J, and Doroshow. Cellminer: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer Research*, 72(14):3499–511, 2012.

[35] Boyd.S, Parikh.N, Chu.E, Peleato.B, and Eckstein.J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[36] Favaro.P, Vidal.R, and Ravichandran.A. A closed form solution to robust subspace estimation and clustering. 42(7):1801–1807, 2011.

[37] Kumar.A, Rai.P, and Daum.H. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, 2011.

[38] Seiva.F.R, Amauchi.J.F, Rocha.K.K, Ebaid.G.X, Souza.G, Fernandes.A.A, Cataneo.A.C, and Novelli.E.L. Alcoholism and alcohol abstinence: N-acetylcysteine to improve energy expenditure, myocardial oxidative stress, and energy metabolism in alcoholic heart disease. *Alcohol*, 43(8):649–56, 2009.

[39] Diniz.Y.S, Rocha.K.K, Souza.G.A, Galhardi.C.M, G.M Ebaid, Rodrigues.H.G, Novelli.F.J.L, Cicogna.A.C, and Novelli.E.L. Effects of n-acetylcysteine on sucrose-rich diet-induced hyperglycaemia, dyslipidemia and oxidative stress in rats. *European Journal of Pharmacology*, 543(1-3):151–7, 2006.

[40] Dunkel.L. Use of aromatase inhibitors to increase final height. *Molecular and Cellular Endocrinology*, 254-255(254-255):207–16, 2006.

[41] Mcculley.A, Haarer.B, Viggiano.S, Karchin.J, and Feng.W. Chemical suppression of defects in mitotic spindle assembly, redox control, and sterol biosynthesis by hydroxyurea. *G3 (Bethesda, Md.)*, 4(1):39–48, 2014.

[42] Maute.R.L, Schneider.C, Sumazin.P, Holmes.A, Califano.A, Basso.K, and Dallafavera.R. trna-derived microrna modulates proliferation and the dna damage response and is down-regulated in b cell lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, 110(4):1404, 2013.

[43] Munkley.J, Livermore.K.E, Mcclurg.U.L, Kalna.G, Knight.B, Mccullagh.P, Mcgrath.J, Crundwell.M, Leung.H.Y, and Robson.C.N. The pi3k regulatory subunit gene pik3r1 is under direct control of androgens and repressed in prostate cancer cells. *Oncoscience*, 2(9):755–764, 2014.

[44] Li.N, Dhar.S.S, Chen.T, Kan.P, Wei.Y, Kim.J, Chan.C, Lin.H, Hung.M, and Lee.M.G. Jarid1d is a suppressor and prognostic marker of prostate cancer invasion and metastasis. *Cancer Research*, 76(4):831, 2016.

[45] Yang.F, Friedrichs.W.E, Cupples.R.L, Bonifacio.M.J, Sanford.J.A, Horton.W.A, and Bowman.B.H. Human ceruloplasmin. tissue-specific expression of transcripts produced by alternative splicing. *Journal of Biological Chemistry*, 265(18):10780–5, 1990.

[46] Mizutani.S, Pauwels.E, Stoven.V, Goto.S, and Yamanishi.Y. Relating drugprotein interaction network with drug side effects. *Bioinformatics*, 28(18):i522–i528, 2011.

[47] Ashburner.M, Ball.C.A, Blake.J.A, Botstein.D, Butler.H, Cherry.J.M, Davis.A.P, Dolinski.K, Dwight.S.S, and Eppig.J.T. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.

Limin Li obtained her Bachelor and Master degrees from Zhejiang University in 2004 and 2006, respectively. She got her Ph.D degree in mathematics at the University of Hong Kong in 2010. She then worked as a postdoctoral fellow in Max Planck Institute of Intelligent System. She is currently an associate professor at School of Mathematics and Statistics in Xi'an Jiaotong University, Xi'an, China. Her research interests include machine learning and the applications in bioinformatics.

Menglan Cai obtained her Bachelor degree from Fujian Normal University in 2015. She is currently a master student at School of Mathematics and Statistics in Xi'an Jiaotong University, Xi'an, China. Her research interest is machine learning.